IEEE Access
Multidisciplinary | Rapid Review | Open Access Journal

Received January 8, 2021, accepted January 20, 2021, date of publication January 25, 2021, date of current version February 2, 2021.

Digital Object Identifier 10.1109/ACCESS.2021.3054332

Facial Expression Recognition Using Hybrid Features of Pixel and Geometry

CHANG LIU[1,2], KAORU HIROTA[1,2], (Life Member, IEEE), JUNJIE MA[3], ZHIYANG JIA[1,2], (Member, IEEE), AND YAPING DAI[1,2]

[1]School of Automation, Beijing Institute of Technology, Beijing 100081, China
[2]State Key Laboratory of Intelligent Control and Decision of Complex Systems, Beijing Institute of Technology, Beijing 100081, China
[3]Department of Computer Science and Technology, Tsinghua University, Beijing 100084, China

Corresponding author: Yaping Dai (daiyaping@bit.edu.cn)

This work was supported in part by the National Talents Foundation of China under Grant No. WQ20141100198, in part by the Beijing Municipal Natural Science Foundation under Grant No. 3192028, and in part by the Beijing Municipal Natural Science Foundation under Grant No. L191020.

ABSTRACT Facial Expression Recognition (FER) has long been a challenging task in the field of computer vision. Most of the existing FER methods extract facial features on the basis of face pixels, ignoring the relative geometric position dependencies of facial landmark points. This article presents a hybrid feature extraction network to enhance the discriminative power of emotional features. The proposed network consists of a Spatial Attention Convolutional Neural Network (SACNN) and a series of Long Short-term Memory networks with Attention mechanism (ALSTMs). The SACNN is employed to extract the expressional features from static face images and the ALSTMs is designed to explore the potentials of facial landmarks for expression recognition. A deep geometric feature descriptor is proposed to characterize the relative geometric position correlation of facial landmarks. The landmarks are divided into seven groups to extract deep geometric features, and the attention module in ALSTMs can adaptively estimate the importance of different landmark regions. By jointly combining SACNN and ALSTMs, the hybrid features are obtained for expression recognition. Experiments conducted on three public databases, FER2013, CK+, and JAFFE, demonstrate that the proposed method outperforms the previous methods, with the accuracies of 74.31%, 95.15%, and 98.57%, respectively. The preliminary results of Emotion Understanding Robot System (EURS) indicate that the proposed method has the potential to improve the performance of human-robot interaction.

INDEX TERMS Facial expression recognition, long short-term memory network, relative geometric position dependency, hybrid feature, attention mechanism.

I. INTRODUCTION

Facial expressions, which convey useful nonverbal cues in daily social communication, are one of the most important features for recognizing the emotional states of human beings. Due to its potential applications in a multiple of research fields, such as affective computing [1], computer vision [2], medical assessment [3], and Human-Robot Interaction (HRI) [4], Facial Expression Recognition (FER) has drawn an upsurge of interest in recent years. Numerous studies have been conducted on emotion recognition problems in facial expression images during the last decades. However, distinguishing facial expressions accurately remains a challenging task because the irrelevant facial information impacts the FER performance. The irrelevant information comes from variant poses, partial occlusion (e.g. hair, glasses), and background clutter. Capturing and representing the most discriminative expression-related features is a key issue to be addressed in facial expression analysis.

Traditional methods attempt to improve the discrimination of expressional features through human design and selection. Facial Action Coding System (FACS) [5] represents facial characteristics by measuring the observable movements of Action Units (AUs), but the changes on the face are sometimes subtle and difficult to be detected. A multitude of other traditional methods, such as Local Binary Patterns (LBP) [6], Histogram of Oriented Gradient (HOG) [7], Gabor wavelet [8], and Scale-invariant Feature Transform (SIFT) [9], address the FER problem by capturing local features. However, these methods mainly focus on the extraction

The associate editor coordinating the review of this manuscript and approving it for publication was Jiankang Zhang.

of shallow features that are relatively singular, resulting in low recognition accuracy.

Compared to traditional methods, the deep learning-based ones have more potential in deep feature extraction. In the last decade, many deep neural network architectures, such as Convolutional Neural Network (CNN) [10], Deep Belief Network (DBN) [11], Generative Adversarial Network (GAN) [12], are successfully applied to FER task, and achieve the state-of-the-art recognition results. Moreover, deep learning-based methods show robust generalization ability. Despite the achievements of the deep learning-based methods, most of them are only designed to extract expressional features from facial pixel images without exploring the relationship between the relative geometric position features of facial landmarks and expressions [13].

Therefore, this article proposes a hybrid feature extraction network to improve the discriminative power of expression-related features. The hybrid features consider two kinds of facial features: 1) pixel-level features, and 2) deep geometry-level features. The proposed model uses a Spatial Attention CNN (SACNN) to extract the pixel-level features based on face pixels and employs a series of Long Short-term Memory (LSTM) networks with Attention mechanism (ALSTMs) to learn the relative geometric position dependencies of facial landmarks. We propose a feature descriptor to characterize the relative geometric position correlation of facial landmarks in a deep-learning way, which is referred to as the deep geometric feature. Motivated by the fact that different facial regions contribute unequally to expressions, a grouping strategy is introduced to divide the facial landmarks into seven groups for local deep geometric feature learning. The attention mechanism is implemented to learn weight vectors and recalibrate the the weights to local geometric features adaptively. The holistic geometric features which integrate all local features are further combined with the pixel-level features to form the hybrid features for expression classification. The contributions of this work are summarized as follows:

- A hybrid feature extraction network that combines the pixel-level feature and the deep geometric feature is developed to tackle facial expression recognition.
- A deep geometric feature descriptor that characterizes the relative geometric position dependencies of facial landmarks is proposed to exploit the facial landmarks for expression recognition.
- A landmark grouping strategy is introduced to divide the facial landmarks which are sent to the ALSTMs network to learn the local-holistic geometric features.
- The experimental results on the public expression databases demonstrate that the proposed method outperforms the previous methods.

The remainder of this article is structured as follows. Section II briefly introduces the related work on FER. The details of the proposed method are presented in Section III. Section IV compares and discusses the experimental results, and Section V shows the preliminary results of a practical emotion understanding robot system. The conclusions of this work are given in Section VI.

TABLE 1. Recent Research Methods.

Expression recognition	Method example
Face detection	Dlib: Amos *et al.* [14]
	Viola-Jones: Viola *et al.* [15]
	MTCNN: Zhang *et al.* [16]
Feature extraction	Handcrafted feature: Shan *et al.* [17] [18] [19]
	GAN: Zhang *et al.* [12]
	DBN: Liu *et al.* [21]
	DNN: Jain *et al.* [22]
	CNN: Ng *et al.* [23] [25] [26]
	ECNN: Jung *et al.* [27] [28]
	CNN + Handcrafted feature: Levi *et al.* [29] [30]
	CNN-LSTM: Yu *et al.* [31] [32]
Attention	Attention-CNN: Gan *et al.* [38] [39] [40] [41]

II. RELATED WORK

In this section, we present the related work on FER system, the expressional feature extraction using deep learning, and the attention mechanism. Table 1 summarizes the different approaches used by the recent FER studies.

A. FACIAL EXPRESSION RECOGNITION SYSTEM

In general, a FER system consists of three components: face detection, feature extraction, and expression classification. In face detection, face detectors such as Dlib [14], Viola-Jones [15], and MTCNN [16] are used to locate and crop faces from complex backgrounds. The feature extraction aims at capturing facial features that are related to expressions. The features are grouped into handcrafted features and learning-based features. Handcrafted features usually require human design and selection elaborately [17], [18], [19]. Learning-based features refer to the high-level abstractions extracted with deep learning techniques. Compared to handcrafted features, learning-based features are more robust to face position changes and scale variations in FER tasks [20]. Hybrid features, which enrich the expression representation, are a combination of two or more features. After the feature extraction, features are passed to a classifier, such as Support Vector Machines (SVM), Random Forest (RF), or softmax loss layer, to predict the expression category to which the given face belongs.

B. FEATURE EXTRACTION USING DEEP LEARNING

Due to the ability of extracting deep-level semantic features and the outstanding recognition performance, deep learning methods are applied to facial expression analysis [12], [21]–[23]. As CNN performs better than other deep learning methods, CNN architectures are widely employed to conduct feature extraction and recognition [24]. Mollahosseini *et al.* [25] present a deep network based on inception structures to increase the depth of the network for FER, which shows a good performance. Arriaga *et al.* [26] use global average pooling to remove the fully connected layers and derive a reduced model for real-time FER.

To improve the discriminability of facial features, some researches combine different CNNs as an ensemble model to conduct expression recognition. A joint fine-tuning method is presented by Jung *et al.* [27] to integrate two separate CNNs and fuse the facial features by weighted summation. Yu *et al.* [28] propose an expression recognition model which employs three different CNNs to complement each other to obtain richer features. However, Ensemble Convolutional Neural Networks (ECNNs) require intensive computing resources and training time.

Several works attempt to combine handcrafted feature extraction techniques with deep learning networks to enrich facial features for FER. For example, Levi *et al.* [29] propose a mapped LBP feature for illumination-invariant FER. The original image and the mapped LBP image are used to train an ECNN to predict expressions. Connie *et al.* [30] propose a hybrid CNN with dense SIFT aggregator for expression recognition, and they achieve outstanding results.

Noting that temporal relations among image sequences are of importance for emotion analysis. More focus has been transferred to algorithms extracting spatial-temporal features simultaneously. As a special form of the recurrent neural network, LSTM is capable of capturing long-term dependencies of sequences with arbitrary lengths. Typically, LSTM takes the features extracted by CNN as input to capture temporal information of facial expressions. The CNN-LSTM framework becomes a preferable choice in video-based FER task [31], [32].

C. ATTENTION MECHANISM

The attention mechanism is widely used in computer vision tasks to address the weakness of convolutions [33], [34]. As for the FER task, some works [35]–[37] focus on learning features from local facial regions and provide a prior knowledge that most of the expressional clues that are beneficial to expression analysis come from the salient facial regions. To highlight the expression-related features, attention mechanism is introduced to adaptively emphasize the important information of facial expression while suppressing the irrelevant information [38], [39]. Li *et al.* [40] propose patch-gated CNN for FER. They use facial landmarks to extract small patches of interest and embed patch-gated units to learn the weights for these patches to obtain region-level attention. Wang *et al.* [41] develop a region attention network to adaptively capture the importance of facial regions for occlusion and pose variant FER.

D. OUR WORK

The above works demonstrate that the feature extraction based on facial pixels and the CNNs with attention mechanism can achieve a good recognition performance. However, there is currently no works exploring how to learn the relative geometric position correlation of facial landmarks for facial expression. This article presents a hybrid feature extraction network for FER. The proposed method uses SACNN to extract the pixel-level facial feature and employs ALSTMs to explore the deep geometric position correlation of facial landmarks. The facial landmarks are divided into seven groups for local-holistic geometric feature extraction and the attention mechanism is utilized to estimate the importance of different landmark regions. The proposed method focuses more on the facial feature extraction on the basis of facial landmarks, helping the network extract more discriminative features that are conducive to recognize expressions.

III. APPROACH FOR HYBRID FEATURE EXTRACTION

In this section, we start by representing the overview of the proposed model, then introduce the spatial attention CNN for pixel feature extraction, the detection and grouping strategy of facial landmarks, and the attention-LSTMs model.

A. OVERVIEW OF SPATIAL ATTENTION CNN ATTENTION LSTMS

The flowchart of the proposed model is illustrated in Fig.1. Image preprocessing is carried out before facial feature extraction. There are two separate branches for feature extraction: one is the pixel-level feature extraction for original face images and the other is the geometric feature extraction for facial landmarks. The former branch uses SACNN architecture to learn facial representations and the spatial attention is embedded to increase the weights of useful features, which makes the network focus more on the expression-related feature. The latter branch consists of a landmark detection module and a series of LSTM networks with attention for exploiting facial landmarks. The landmark detection module is employed to locate the facial landmark points which are further sent to the well-designed ALSTMs to learn deep geometry-level features.

In addressing facial landmarks, a deep feature descriptor is proposed to characterize the relative geometric position features. We introduce a grouping strategy to group the landmarks according to their positions and use the attention mechanism to estimate the importance of the local deep geometric features of grouped landmarks. Then the pixel-level feature and the geometry-level feature are fused as the hybrid feature through concatenate operation for classification. At last, the classification loss is calculated at the last fully connected layer, and the overall network is optimized by minimizing the loss function.

B. SPATIAL ATTENTION CNN FOR PIXEL-LEVEL FEATURE EXTRACTION

CNN is extensively used in the recognition of facial expressions because of the ability of capturing deep level feature abstractions. The specific CNN architecture in the proposed network for pixel-level feature extraction is inspired by the VGG-Net, and a spatial attention module is used to VGG19 network as the pixel feature extractor, shown as Fig.2. The details of the pixel feature extractor are shown in Table 2. The size of the input image is 44×44. Conv($ks \times ks \times c_{\text{in}} \times c_{\text{out}}$) stands for the convolutional layer with the kernel size $ks \times ks$, where c_{in} and c_{out} represent the

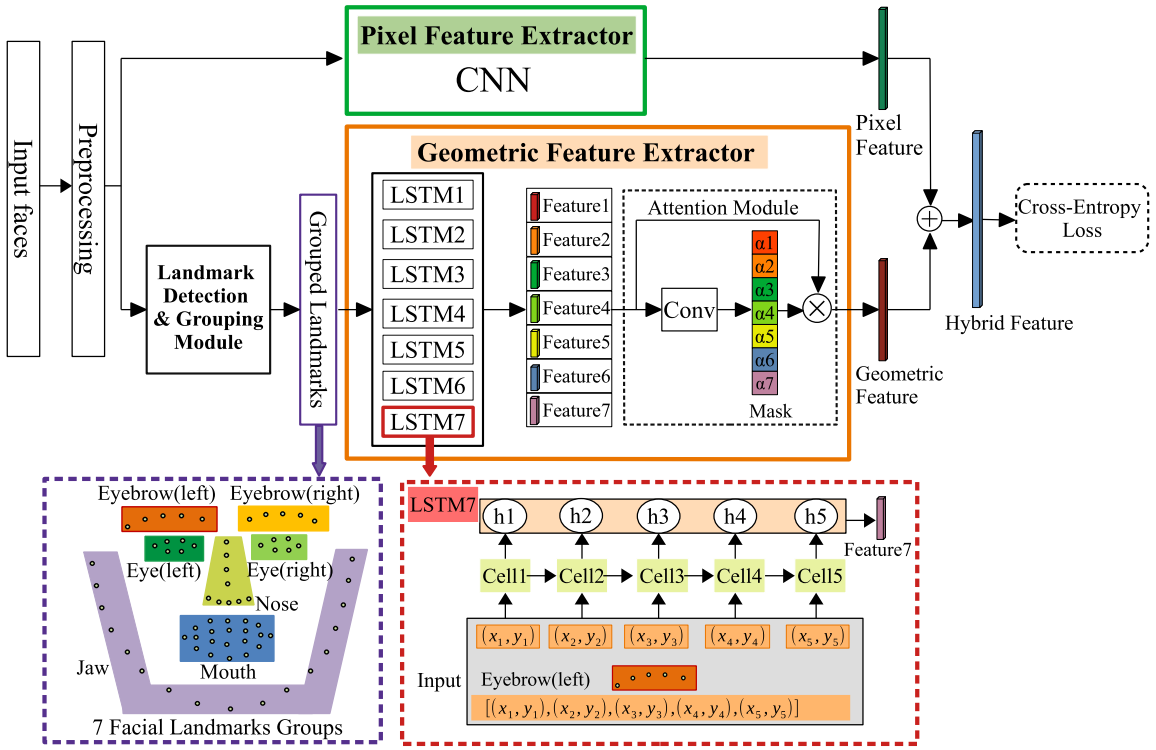


FIGURE 1. An overview of the proposed network, which contains two separate feature extractors. The first one (in the green rectangle) is the SACNN that extracts the pixel-level feature based on facial images. The second one (in the orange rectangle) is the ALSTMs for geometry-level feature extraction based on the facial landmarks in different facial regions. For instance, the LSTM in the red dashed rectangle takes the landmark coordinates of the left eyebrow to extract the local geometric feature. The detected facial landmarks and the corresponding groups are shown in the purple dashed rectangle.

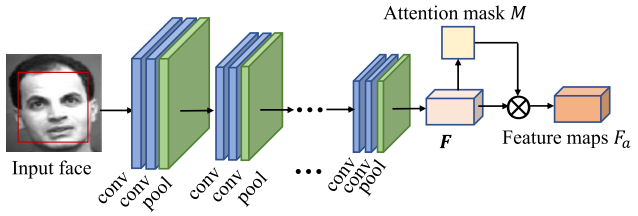


FIGURE 2. The CNN architecture in the proposed model.

number of input channels and output channels, respectively. MaxPool($k \times k$) denotes a $k \times k$ max pooling layer. Batch Normalization (BN) layer is added after each convolution block to reduce the internal covariate shift, and Rectified Linear Unit (ReLU) is used for activation. A spatial attention layer is employed to reduce the irrelevant information by estimating a scaler weight that denotes the importance of the corresponding group of pixels. It is expected that the irrelated regions of facial images are assigned low importance weights. The extracted CNN features are fed into the attention network, which outputs an attentive mask to quantify the importance of each position in feature maps. The extracted features are then weighted by the attentive mask. Mathematically, we denote the extracted feature maps as $\boldsymbol{h}_c$ and the attentive mask as M_c. A one-layer convolutional model is adopted to obtain the attention mask, which is formulated as follows:

$$\boldsymbol{M}_c = f_a(\boldsymbol{W}_c * \boldsymbol{h}_c + \boldsymbol{B}_c) \tag{1}$$

TABLE 2. The Network Configuration of CNN.

Input:	$44 \times 44 \times 3$	Output size
VGG-Net	Conv($3 \times 3 \times 3 \times 64$), BN, ReLU	$(44 \times 44 \times 64)$
	Conv($3 \times 3 \times 64 \times 64$), BN, ReLU	$(44 \times 44 \times 64)$
	MaxPool(2×2)	$(22 \times 22 \times 64)$
	Conv($3 \times 3 \times 64 \times 128$), BN, ReLU	$(22 \times 22 \times 128)$
	Conv($3 \times 3 \times 128 \times 128$), BN, ReLU	$(22 \times 22 \times 128)$
	MaxPool(2×2)	$(11 \times 11 \times 128)$
	Conv($3 \times 3 \times 128 \times 256$), BN, ReLU	$(11 \times 11 \times 256)$
	Conv($3 \times 3 \times 256 \times 256$), BN, ReLU$\times 2$	$(11 \times 11 \times 256)$
	MaxPool(2×2)	$(5 \times 5 \times 256)$
	Conv($3 \times 3 \times 256 \times 512$), BN, ReLU	$(5 \times 5 \times 512)$
	Conv($3 \times 3 \times 512 \times 512$), BN, ReLU$\times 3$	$(5 \times 5 \times 512)$
	MaxPool(2×2)	$(2 \times 2 \times 512)$
	Conv($3 \times 3 \times 512 \times 512$), BN, ReLU$\times 4$	$(2 \times 2 \times 512)$
Attention	Conv($1 \times 1 \times 512 \times 1$), tanh	(2×2)
	FC-layer(512)	(512)

where $\boldsymbol{W}_c$ and $\boldsymbol{B}_c$ are respectively the weights and bias of the convolutional layer, and $\boldsymbol{B}_c$ is randomly initialized. $f_a(\cdot)$ is the tanh activation function, and the value of the attentive mask can be limited within the range of $(-1, 1)$. The features with positive weights are regarded as expression-related features while the features with negative weights are regarded as redundant features which should be filtered out. The importance of corresponding pixel can be reweighted, and the weighted pixel-level feature map $\boldsymbol{h}_p$ is obtained

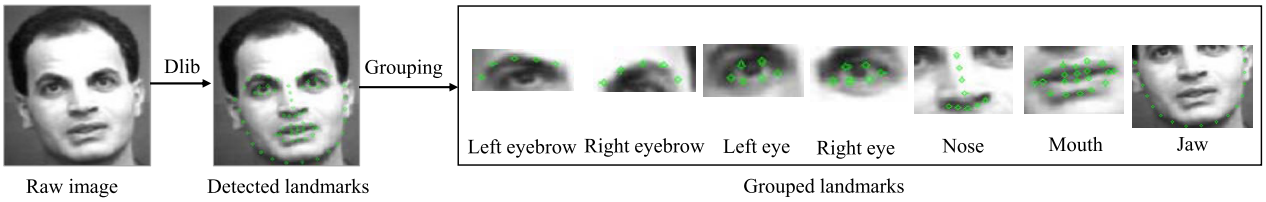


FIGURE 3. The detection and grouping of 68 facial landmarks. The 68 detected facial landmarks are divided into seven components, i.e., left eyebrow, right eyebrow, left eye, right eye, nose, mouth, and jaw.

TABLE 3. Component Partition of 68 Facial Points.

Facial Region	Landmark Coordinates	Facial Feature Vector
Left eyebrow	$\{(x_1^k, y_1^k)\}, k = 1, 2, \ldots, 5$	$\boldsymbol{v}_1 = (x_1^1, y_1^1, x_1^2, y_1^2, \ldots, x_1^5, y_1^5)$
Right eyebrow	$\{(x_2^k, y_2^k)\}, k = 1, 2, \ldots, 5$	$\boldsymbol{v}_2 = (x_2^1, y_2^1, x_2^2, y_2^2, \ldots, x_2^5, y_2^5)$
Left eye	$\{(x_3^k, y_3^k)\}, k = 1, 2, \ldots, 6$	$\boldsymbol{v}_3 = (x_3^1, y_3^1, x_3^2, y_3^2, \ldots, x_3^6, y_3^6)$
Right eye	$\{(x_4^k, y_4^k)\}, k = 1, 2, \ldots, 6$	$\boldsymbol{v}_4 = (x_4^1, y_4^1, x_4^2, y_4^2, \ldots, x_4^6, y_4^6)$
Nose	$\{(x_5^k, y_5^k)\}, k = 1, 2, \ldots, 9$	$\boldsymbol{v}_5 = (x_5^1, y_5^1, x_5^2, y_5^2, \ldots, x_5^9, y_5^9)$
Mouth	$\{(x_6^k, y_6^k)\}, k = 1, 2, \ldots, 20$	$\boldsymbol{v}_6 = (x_6^1, y_6^1, x_6^2, y_6^2, \ldots, x_6^{20}, y_6^{20})$
Jaw	$\{(x_7^k, y_7^k)\}, k = 1, 2, \ldots, 17$	$\boldsymbol{v}_7 = (x_7^1, y_7^1, x_7^2, y_7^2, \ldots, x_7^{17}, y_7^{17})$

by

$$\boldsymbol{h}_p = \boldsymbol{M}_c \odot \boldsymbol{h}_c \tag{2}$$

where the symbol "$\odot$" represents element-wise product.

C. FACIAL LANDMARK DETECTION AND GROUPING STRATEGY

The locations of the fiducial facial landmark points around the facial components and facial contour capture the rigid and non-rigid deformations caused by facial expressions. The facial landmark detection is hence important for expression recognition. Geometric facial features for expression analysis are based on locating the landmarks and determining the relative position relation of associated facial components, i.e., eyebrows, eyes, nose, mouth, and jaw. Shahid *et al.* [42] propose local facial shape harmonic features to recognize expressions on the basis of eleven sub-local facial regions.

In this article, the Dlib toolkit is adopted to detect the position of 68 facial landmarks. Each of the landmarks is represented by a 2-dimensional Cartesian coordinate as (x, y). Fasel *et al.* [20] point out that facial expressions are caused by the changes in facial behavior and are closely related to some specific areas rather than the whole face. According to the abovemetioned principles, this article introduces a grouping strategy to divide the facial landmarks into seven groups in terms of their different positions on the face, including left eyebrow, right eyebrow, left eye, right eye, nose, mouth, and jaw. The detection and grouping of facial landmarks are illustrated in Fig.3. For example, there are 5 landmark points in the left eyebrow area, and their corresponding 2-dimensional coordinates $\{(x_1^k, y_1^k)\}, k = 1, 2, \ldots, 5$, work together to form a 10-dimensional feature vector $\boldsymbol{v}_1 = (x_1^1, y_1^1, x_1^2, y_1^2, \ldots, x_1^5, y_1^5)$. For the coordinate (x_j^i, y_j^i) of a landmark point, the subscript denotes the j-th component and the superscript represents the i-th landmark. Similarly, the rest groups of facial landmarks can be reprented, and their corresponding feature vectors are listed in Table 3.

D. ATTENTION LSTMs FOR GEOMETRY-LEVEL FEATURE EXTRACTION

To capture the latent information hidden in the landmarks of different facial regions, ALSTMs is designed to extract the deep geometry-level representations from the Cartesian coordinates of the grouped facial landmarks. Specifically, seven LSTMs take the corresponding seven facial landmark sequences $\boldsymbol{v}_k$, $(k = 1, 2, \ldots, 7)$ and map the input sequences into output sequences separately by calculating the activations of the cell units in the network. As an example, the left eyebrows landmark vector $\boldsymbol{v}_1 = (x_1^1, y_1^1, x_1^2, y_1^2, \ldots, x_1^5, y_1^5)$ can be regarded as five sequences, each of which is denoted as a 2-dimensional Cartesian coordinate. The landmark sequences are passed to the corresponding LSTM to capture the relative positional dependencies, then the local deep geometric feature of the left eyebrows is obtained. The process of extracting the geometric feature is represented by the following formulations recursively:

$$\boldsymbol{i}_t = \sigma(\boldsymbol{W}_{vi}\boldsymbol{v}_t + \boldsymbol{W}_{hi}\boldsymbol{h}_{t-1} + \boldsymbol{W}_{ci}\boldsymbol{c}_{t-1} + \boldsymbol{b}_i) \tag{3}$$

$$\boldsymbol{f}_i = \sigma(\boldsymbol{W}_{vf}\boldsymbol{v}_t + \boldsymbol{W}_{hf}\boldsymbol{h}_{t-1} + \boldsymbol{W}_{cf}\boldsymbol{c}_{t-1} + \boldsymbol{b}_f) \tag{4}$$

$$\boldsymbol{c}_t = \boldsymbol{f}_t\boldsymbol{c}_{t-1} + \boldsymbol{i}_t \tanh(\boldsymbol{W}_{vc}\boldsymbol{v}_t + \boldsymbol{W}_{hc}\boldsymbol{h}_{t-1} + \boldsymbol{b}_c) \tag{5}$$

$$\boldsymbol{o}_t = \sigma(\boldsymbol{W}_{xo}\boldsymbol{v}_t + \boldsymbol{W}_{ho}\boldsymbol{h}_{t-1} + \boldsymbol{W}_{co}\boldsymbol{c}_t + \boldsymbol{b}_o) \tag{6}$$

$$\boldsymbol{h}_t = \tanh(\boldsymbol{c}_t) \tag{7}$$

where $\boldsymbol{i}_t, \boldsymbol{f}_t, \boldsymbol{c}_t$, and $\boldsymbol{o}_t$ are the activation vectors of the input gate, forget gate, memory cell, and output gate in the LSTM model, respectively. $\boldsymbol{v}_t$ and $\boldsymbol{h}_t$ are separately the input and hidden vectors at the t-th time step. $\boldsymbol{W}_{\alpha\beta}$ denotes the weights

matrix between α and β. $\boldsymbol{b}_\alpha$ is the bias of α, and $\sigma(\cdot)$ represents the sigmoid function $\sigma(x) = 1/(1 + e^{-x})$. It is noted that the cell number of each LSTM is equal to the number of the input facial landmarks. Likewise, the deep geometry-level features of the other facial regions are extracted through the corresponding LSTMs. Then the holistic deep geometry-level features $\boldsymbol{h}_l$ of the whole face are obtained by concatenating all the separately learned features $\boldsymbol{h}_i$ of all seven facial regions, as follows:

$$\boldsymbol{h}_l = [\boldsymbol{h}_1, \boldsymbol{h}_2, \ldots, \boldsymbol{h}_7]. \tag{8}$$

The FACS points out human expressions can be represented by a set of AUs and reveals that expressions are relevant to a few facial regions. Different facial regions contribute unequally to expression recognition, which motivates us to treat the landmarks in different facial regions discriminatively in terms of their importance for the classification. An attention module is employed to highlight discriminative features learned from the expression-related landmarks. In particular, the extracted local deep geometric features of all seven LSTMs are fed into the attention network, which outputs an attentive mask to quantify the importance of the geometric features. The extracted features are weighted by the attentive mask and the holistic geometry-level features are obtained. Concretely, we denote the extracted features as $\boldsymbol{h}_l$ and the attentive mask as $\boldsymbol{M}_l$. An one-layer convolutional model is used to get the attentive mask, which is formulated as follows:

$$\boldsymbol{M}_l = f_a(\boldsymbol{W}_g * \boldsymbol{h}_l + \boldsymbol{B}_g) \tag{9}$$

where $\boldsymbol{W}_g$ is the convolutional kernels of the attention network, and $\boldsymbol{B}_g$ is the corresponding bias. The symbol "*" indicates the operation of convolution and $f_a(\cdot)$ is softmax activation function. In our model, the kernel size of $\boldsymbol{W}_a$ is 1×1 and the bias $\boldsymbol{B}_a$ is randomly initialized. Each weight of the attentive mask is only related to the features in the corresponding position. Therefore, each attentive weight can reflect the degree of the importance of different facial landmarks. To obtain the features beneficial to the expression classification, the extracted features are weighted by the attentive mask as follows:

$$\boldsymbol{h}_g = \boldsymbol{M}_l \odot \boldsymbol{h}_l \tag{10}$$

where "$\odot$" is the operation of element-wise multiplication and $\boldsymbol{h}_g$ indicates the final geometry-level feature.

E. LOSS FUNCTION

The loss function is required to evaluate the training of the model in the weights update. The softmax activation is used for classification in the last layer of fully connected layers. Denote the ground-truth expression label for the k-th sample as a one-hot vector $\boldsymbol{y}_k = [y_k^1, y_k^2, \ldots, y_k^c, \ldots, y_k^C]$, where C is the number of categories. The proposed model is trained under the supervision of cross-entropy loss, which is defined as follows:

$$Loss_{CE} = -\frac{1}{K}\sum_{k=1}^{K}\sum_{c=1}^{C} y_k^c \log p_k^c \tag{11}$$

where p_k^c represents the predicted probability of the c-th category for the k-th sample.

IV. EXPERIMENTS ON EXPRESSION RECOGNITON

In this section, we introduce the detailed evaluation of the proposed method, including the database description, implementation details of experiments, and the experimental results.

A. DATABASE DESCRIPTION

To evaluate the proposed method, we conduct the experiments on three public facial expression recognition databases: the FER2013 database [43], the extended Cohn-Kanade (CK+) database [44], and the Japanese Female Facial Expression (JAFFE) database [45]. These databases cover different scales of face images and challenging conditions.

1) ***FER2013:*** The FER2013 database is introduced during the ICML 2013 Challenges. FER2013 is a large-scale and unconstrained database collected automatically by the Google image search API. All images have been registered and resized to 48×48 grayscale images of faces after adjusting the cropped region. FER2013 contains 28,709 training images, 3,589 validation images, and 3,589 test images with seven basic expression labels (anger, disgust, fear, happiness, sad, surprise, and neutral). The samples of the FER2013 database are shown in Fig.4.

2) ***CK+:*** The CK+ database is the most extensively used laboratory-controlled database for evaluating FER systems. CK+ contains 593 video sequences from 123 subjects. The image sequences vary in duration from 10 to 60 frames and show a shift from the neutral facial expression to the peak expression. Among these videos, 327 sequences from 118 subjects are labeled with seven basic expression labels (anger, contempt, disgust, fear, happiness, sad, and surprise) based on the FACS. The samples of the CK+ database are shown in Fig.5.

3) ***JAFFE:*** The JAFFE database holds 213 images in total of seven facial expressions (anger, disgust, fear, happy, sad, surprise, and neutral) posed by ten Japanese female participants. The images are in size of 256×256 pixels and the expresser have 2-4 samples for every expression. The samples of the JAFFE database are shown in Fig.6. The number of images for each prototypical expression in the databases are listed in Table 4.

Preprocessing: The image resolutions of the FER2013 database, the CK+ database, and the JAFFE database are 48×48, 640×490, and 224×224, respectively. The Dlib toolkit is used for face detection, and the images that identify faces are preserved. As the input dimension of a deep network is fixed, we select the size of 48×48 as the final image resolution. For the CK+ database, the last three frames with peak expressions of face sequences are selected for experiments.

FIGURE 4. Some samples of FER2013 database.

FIGURE 5. Some samples of CK+ database.

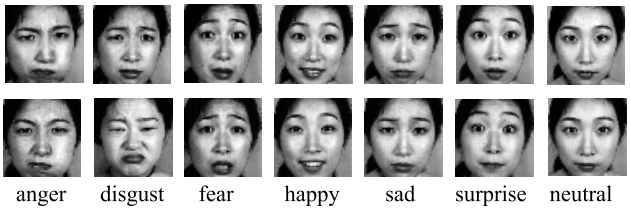


FIGURE 6. Some samples of JAFFE database.

TABLE 4. Number of Images for Each Expression in Databases. AN, DI, FE, HA, SA, SU, Ne, and Co Stand for Anger, Disgust, Fear, Happiness, Sad, Surprised, Neutral, and Contempt, Respectively.

Database	AN	DI	FE	HA	SA	SU	NE	CO
FER2013	4953	547	5121	8989	6077	4002	6198	-
CK+	45	59	25	69	28	83	-	17
JAFFE	30	29	32	31	31	30	30	-

The key face area of each image is detected and cropped from the raw face image. The cropped image is resized to the resolution of 48×48.

B. IMPLEMENTATION DETAILS

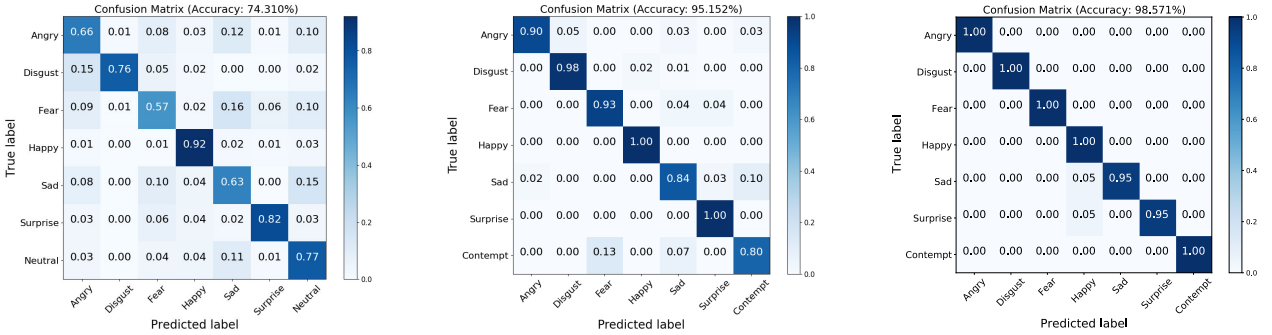
The proposed method is implemented using PyTorch deep learning framework on a GTX 1080Ti with 11 GB machine. The optimization method used for weights update is the Stochastic Gradient Descent (SGD) algorithm, where the learning rate starts at 0.01. The momentum, weight decay, and batch size are 0.9, 0.0005, and 32, respectively. Dropout is adopted in the CNN to avoid overfitting, and the dropout ratio is set to 0.5. The number of hidden layers and hidden layer nodes in the LSTM are 2 and 128, respectively. For the FER2013 database, the total number of epochs is set to 200. The learning rate begins to decrease after 30 epochs and is decreased by multiplying it with 0.98 after each epoch. The number of epochs for the CK+ database and the JAFFE database is 150. The learning rate begins to decrease after 30 epochs and is decreased by multiplying it with 0.95 after each 3 epochs. In the training stage, we randomly cut five images with the size of 44×44 from each training sample as the training data and feed them to the CNN to learn pixel-level features. During the testing phase, an integrated approach is used to reduce outliers. Five images with the same size of training images are cropped for validation in the upper left, lower left, upper right, lower right, and center. The average probabilities based on the five cropped face images are obtained, and the maximum output classification is the prediction of the corresponding expression.

C. RESULTS ON THE DATABASES

Experiments are conducted on FER2013, CK+, and JAFFE to evaluate the performance of the proposed model. VGG19 with spatial attention is selected as the baseline model, and the pixel-level feature extractor in the proposed model adopts the same convolutional structure but removing the last softmax layer. To analyze the effect of the ALSTMs in extracting facial features from the grouped landmarks, two models with different geometry-level feature extractors are compared. The former is the proposed SACNN-ALSTMs model which uses ALSTMs to extract the geometric feature from facial landmarks. In the latter SACNN-LSTM model, we replace the ALSTMs with one LSTM which takes all facial landmarks of the whole face as the input for geometric feature extraction. Therefore, the contrast experiments are carried out on the baseline model, the SACNN-LSTM model, and the SACNN-ALSTMs model.

1) Results on FER2013: Experiments are conducted on the unconstrained FER2013 database to evaluate the performance of the proposed model when dealing with complex variations. As shown in Table 5, the accuracies of the baseline model, the SACNN-LSTM, and the SACNN-ALSTMs achieve 71.36%, 73.22%, and 74.31%, respectively. The average recognition accuracy of the SACNN-ALSTMs model and the SACNN-LSTM model outperform the baseline model by 2.95% and 1.86%, respectively. The improvement in accuracy is caused by the geometry-level feature extracted from facial landmarks, which proves the geometric feature extracted from facial landmarks are beneficial to improve the expression recognition performance. The accuracy of the proposed SACNN-ALSTMs model outperforms the SACNN-LSTM model by 1.09%, which indicates that the landmark grouping strategy for learning features from the landmarks of different facial regions is more effective. It may be explained that the attention mechanism assigns larger weights to relative positional dependencies of the facial landmarks in the areas associated with facial expressions.

The detailed recognition accuracies (confusion matrix) of each facial expression are shown in Fig.7(a). In addition to the confusion matrix, we compute precision, recall, and F1-score to further measure the expression recognition performance of the proposed model. The experimental results are presented in Table 6. The Receiver Operating Curve (ROC) of the proposed network on the FER2013 database is shown in Fig.8, from which we can observe that the proposed model achieves



(a) FER2013 (b) CK+ (c) JAFFE

FIGURE 7. The confusion matrices of the FER2013 database, the CK+ database, and the JAFFE database.

TABLE 5. The Accuracy (%) Results on Three Databases.

Model	FER2013	CK+	JAFFE
Baseline model	71.36	92.32	96.43
SACNN-LSTM (ours)	73.22	94.04	97.86
SACNN-ALSTMs (ours)	**74.31**	**95.15**	**98.57**

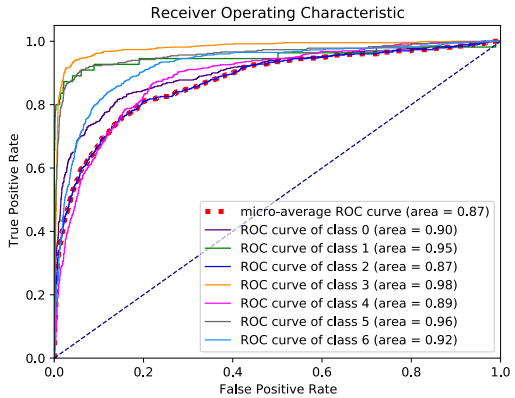


FIGURE 8. The ROC of the SACNN-ALSTMs model on FER2013 dataset. The red dotted line represents the average ROC curve, and the solid lines are the ROC curves of seven categories. Class 0,1,2,3,4,5,6 denote anger, disgust, fear, happy, sad, surprise and neutral, respectively.

good performance with an average Area Under the Curve (AUC) of 0.87. It can be observed from the experimental results that both ''happy'' and ''surprised'' categories are easier to distinguish with high recognition accuracies. It can be explained that these expressions have clear facial muscle movements and shape deformations. The facial expressions of ''fear'', ''sad'', and ''anger'' are easy to be misclassified. The reason is that the expressions of abovementioned three categories are similar to each other, and it is difficult for the proposed model to distinguish them accurately.

2) Results on CK+: To possess more reliable results, 10-fold cross-validation is used for classification on the CK+ database. All the facial images are equally divided into ten groups, nine of which are used for training each time, and the remaining one group is used for the test. Such experiments are repeated ten times, and the average accuracy is taken as the final prediction result.

TABLE 6. Recognition Performance (%) Measure for Each Expression When SACNN-ALSTMs Gives an Average Accuracy of 74.31% on the FER2013 Dataset.

Emotion	Precision	Recall	F1-score
Anger	69.10	65.58	67.29
Disgust	82.35	76.36	79.25
Fear	65.71	56.63	60.83
Happy	89.89	92.04	90.95
Sad	60.88	62.63	61.74
Surprise	86.08	81.73	83.85
Neural	67.93	77.16	72.25

The average recognition accuracies of the baseline model, the SACNN-LSTM model, and the SACNN-ALSTMs model are 92.32%, 94.04%, and 95.15%, respectively, as shown in Table 5. It can be observed from the experimental results that the SACNN-ALSTMs and the SACNN-LSTM outperform the baseline model by 2.83% and 1.72%. The results suggest that the hybrid features are more discriminative for expression classification than the pixel-level features only extracted from the raw face images. This may be owing to the deep geometry-level features learned from facial landmarks offer more latent expression-related information. In addition, the accuracy of the SACNN-ALSTMs model is higher than that of the SACNN-LSTM model by 1.11%, which verifies the effectiveness of the landmark grouping strategy.

The confusion matrix of the predictions of the SACNN-ALSTMs on the CK+ database is shown in Fig.7(b). The experimental results of the precision, recall, and F1-score carried on CK+ database are listed in Table 7. It can be observed from the confusion matrix that ''happy'' and ''surprise'' are recognized with high accuracies for all three models. ''Sad'' and ''contempt'' are relatively hard to distinguish, and they are easily misclassified. This may be due to the lack of training samples in these prototypical facial expressions, which may result in learning more features of other expressions. The ROC curves of the proposed network on the CK+ database are shown in Fig.9, from which it can be observed that the proposed model achieves good performance with an average AUC of 0.97.

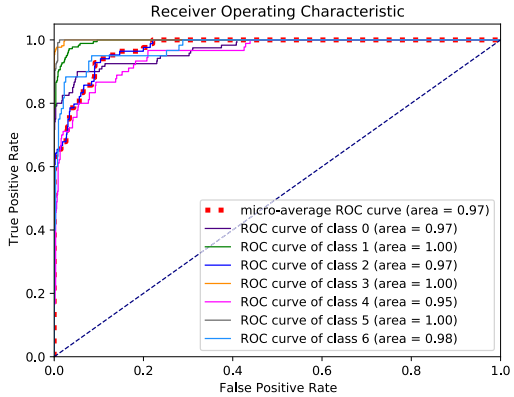


FIGURE 9. The ROC of SACNN-ALSTMs model on CK+ dataset. The red dotted line represents the average ROC curve, and the solid lines are the ROC curves of seven categories. Class 0,1,2,3,4,5,6 denote anger, disgust, fear, happy, sad, surprise and contempt, respectively.

TABLE 7. Recognition Performance (%) Measure for Each Expression When SACNN-ALSTMs Gives an Average Accuracy of 95.15% on the CK+ Dataset.

Emotion	Precision	Recall	F1-score
Anger	98.18	90.00	93.91
Disgust	96.70	97.78	97.24
Fear	90.70	92.86	91.76
Happy	98.63	100.00	99.31
Sad	87.36	84.44	85.88
Surprise	97.56	100.00	98.77
Contempt	80.00	80.00	80.00

TABLE 8. Recognition Performance (%) Measure for Each Expression When SACNN-ALSTMs Gives an Average Accuracy of 98.57% on the JAFFE Dataset.

Emotion	Precision	Recall	F1-score
Anger	100.00	100.00	100.00
Disgust	100.00	100.00	100.00
Fear	100.00	100.00	100.00
Happy	90.91	100.00	95.24
Sad	100.00	95.00	97.44
Surprise	100.00	95.00	97.44
Neural	100.00	100.00	100.00

3) Results on JAFFE: The total number of samples in the JAFFE database is small, but there is a similar number of images showing each expression. Similar to the CK+ database experiments, 10-fold cross-validation is applied for classification on the JAFFE database. As listed in Table 5, the SACNN-ALSTMs model achieves an average recognition accuracy of 98.57%, which outperforms the baseline model and the SACNN-LSTM model by 2.14% and 1.43%. The experimental results demonstrate the superiority of the proposed model on the JAFFE database. The confusion matrix of the SACNN-ALSTMs on the JAFFE database is shown in Fig.7(c). The details of the precision, recall, and F1-score carried on the JAFFE database are listed in Table 8. It can be found that all seven categories of expressions in the JAFFE database are recognized by the proposed model with high accuracies.

TABLE 9. Performance (%) Comparison on the FER2013 Database.

Method	Accuracy
Resnet18	65.68
VGG19	66.54
Tang 2013 [46]	71.2
Devries *et al.* 2014 [47]	67.21
Jeon *et al.* 2016 [48]	70.74
Guo *et al.* 2016 [49]	71.44
Arriaga *et al.* 2017 [26]	66.00
Munasinghe *et al.* 2017 [50]	71.10
Xie *et al.* 2019 [51]	66.20
Sun *et al.* 2020 [52]	72.5
SACNN-ALSTMs (ours)	**74.31**

D. COMPARISON EXPERIMENT WITH OTHER METHODS

1) Comparison on FER2013: The proposed method is compared with some other existing deep learning-based methods on the FER2013 database. The details of the comparison results are listed in Table 9. Tang [46] uses a linear SVM classifier instead of softmax layer and achieves a recognition accuracy of 71.2%. The method proposed by Devries *et al.* [47] obtains an average accuracy of 67.21%. Jeon *et al.* [48] propose a CNN based on the HOG feature and achieve an average accuracy of 70.74%. Guo *et al.* [49] present a deep learning method termed deep neural network with relativity learning which obtains an average accuracy of 71.44%. The method proposed by Arriaga *et al.* [26] obtains an accuracy of 70.74%. Munasinghe *et al.* [50] propose a sequential-based framework which achieves a recognition accuracy of 71.10%. Xie *et al.* [51] propose a deep multi-path CNN with salient region attention and obtain an average accuracy of 72.5%. Sun *et al.* [52] propose an ROI-Attention vectorized CNN model that can locate expression-related regions and estimate the importance of different image regions, and they achieve an average accuracy of 66.20%.

From Table 9, we can observe that the proposed SACNN-ALSTMs outperforms other competitive methods with an accuracy of 74.31%. Compared with other deep learning-based methods, the proposed model extracts hybrid features that combine pixel-level features learned from face pixels and deep geometry-level features learned from facial landmark coordinates. The hybrid features are more discriminative for expression classification. It may be explained that landmark coordinates provide latent features associated with expressions, and the ALSTMs has the ability to capture the relative geometric position correlation among the expression-related facial landmarks.

2) Comparison on CK+: Comparison experiments with other deep learning-based methods are conducted on the CK+ database. The comparison results are listed in Table 10. Taheri *et al.* [53] propose a dictionary-based approach by decomposing expressions in terms of AUs and achieve an accuracy of 88.53%. Happy *et al.* [35] propose a framework to capture appearance features of salient facial patches and obtain an accuracy of 94.69%. The DTAN model and DTGN

TABLE 10. Performance (%) Comparison on the CK+ Database. (Loso: Leave-One-Suject-Out, 5(8,10)-Fold: 5(8,10)-Fold-Cross Validation).

Method	Accuracy	Validation
Resnet18	86.26	10-fold
VGG19	87.35	10-fold
Taheri *et al.* 2014 [53]	88.52	loso
Happy *et al.* 2014 [35]	94.69	10-fold
(DTAN) Jung *et al.* 2015 [27]	91.44	10-fold
(DTGN) Jung *et al.* 2015 [27]	92.35	10-fold
Mollahosseini *et al.* 2016 [25]	93.20	5-fold
Lopes *et al.* 2017 [54]	92.73	8-fold
Zhang *et al.* 2018 [55]	92.35	10-fold
Cai *et al.* 2018 [56]	94.35	10-fold
Jain *et al.* 2019 [22]	93.24	-
Sun *et al.* 2020 [52]	87.20	-
Shahid *et al.* 2020 [42]	94.90	10-fold
SACNN-ALSTMs (ours)	**95.15**	10-fold

model proposed by Jung *et al.* [27] achieve the classification accuracy of 91.44% and 92.35%, respectively. The deep neural network presented by Mollahosseini *et al.* [25] achieves an average accuracy of 93.20%. Lopes *et al.* [54] use a combination of CNN and specific image pre-processing steps to boost the recognition performance, which achieves an accuracy of 92.73%. The image sequences-based network proposed by Zhang *et al.* [55] integrates the feature learning from both spatial and temporal information, and obtains a recognition result of 92.35%. The island loss is proposed by Cai *et al.* [56] to minimize the intra-class distances of deep features while maximizing inter-class distances, and it achieves an accuracy of 94.35%. Jain *et al.* [22] propose an extended deep neural network which achieves an average accuracy of 93.24%. The method proposed by Sun *et al.* [52] obtains an average accuracy of 87.20% on CK+ database. The proposed model achieves an average recognition accuracy of 95.15% on the CK+ database. It can be observed that the SACNN-ALSTMs model outperforms previous methods, which verifies the effectiveness of the proposed method.

3) Comparison on JAFFE: The proposed method is compared with other methods on the JAFFE database. As shown in Table 11, Happy *et al.* [35] capture appearance features of salient facial patches for FER and obtain an accuracy of 91.80% on the JAFFE database. Lopes *et al.* [54] use a combination of CNN and specific image pre-processing steps to boost the recognition performance, which achieves an accuracy of 94.86%. Jain *et al.* [22] propose an extended deep neural network which achieves an average accuracy of 95.23%. Li *et al.* [57] propose face cropping and image rotation methods for CNN training and obtain an accuracy of 97.18%. The ROI-Attention vectorized CNN model [52] can obtain an average accuracy of 92.00% on seven types of expression recognition. The proposed method outperforms the other methods described above, achieving an average accuracy of 98.57%.

The recognition accuracies of the CK+ database and the JAFFE database are higher than that of the FER2013 database. The reason for reduced performance

TABLE 11. Performance (%) Comparison on the JAFFE Database (10-Fold: 10-Fold-Cross Validation).

Method	Accuracy	Validation
Resnet18	89.90	10-fold
VGG19	91.27	10-fold
Happy *et al.* 2014 [35]	91.80	-
Lopes *et al.* 2017 [54]	94.86	10-fold
Jain *et al.* 2019 [22]	95.23	-
Li *et al.* 2020 [57]	97.18	10-fold
Sun *et al.* 2020 [52]	92.00	-
SACNN-ALSTMs (ours)	**98.57**	10-fold

TABLE 12. Complexity Measure for the SACNN-ALSTMs Model.

Model	Parameters (M)	FLOPs (G)
Baseline	203.02	0.898
SACNN-LSTM	430.63	0.907
SACNN-ALSTMs (ours)	396.76	0.929

shown by FER2013 trained model is a variation of faces in terms of diversity and control conditions. The CK+ database and the JAFFE database are collected in a controlled laboratory environment, and the data are all frontal faces that have few background variations.

E. EVALUATION OF COMPLEXITY AND INFERENCE TIME

The Floating Point Operations (FLOPs) and parameters are used to evaluate the time complexity and space complexity of the proposed model, and the comparison results are listed in Table 12. We can see that the proposed SACNN-ALSTMs model is more complex in terms of parameters and FOLPs because the ALSTMs is introduced into the proposed network to deal with the facial landmarks. The inference time is further investigated on the CK+ dataset. The average inference time per image is obtained on an NVIDIA 1080Ti GPU of Linux system with an Intel(R) i9-7900X CPU @3.30GHz. The average inference time of the baseline model, the SACNN-LSTM model, and the proposed SACNN-ALSTMs model are 1.2ms, 1.5ms, and 2.1ms, respectively. Although the proposed model increases in time complexity and space complexity, the increase of inference time is negligible considering the improvement of the GPU hardware computing power.

F. CROSS-DATABASE EXPERIMENT

To evaluate the generalization ability of the proposed model, we conduct six experiments across different databases. The FER2013 and JAFFE do not contain contempt expression, and the CK+ does not include neutral expressions. Therefore, these expressions are neglected. The network is trained on one database and tested on the two remaining databases, as shown in Table 13. Sun *et al.* [52] train the deep multi-path CNN on the FER2013 and AffectNet databases, and test on the JAFFE database. Xie *et al.* [51] conduct cross-database experiments on the CK+ database and the JAFFE database. From the results, we can observe that the proposed

TABLE 13. Performance on Cross-Database Evaluation (%).

Train	Test	Method	Accuracy
FER2013	CK+	SACNN-ALSTMs(ours)	62.78
	JAFFE	SACNN-ALSTMs(ours)	65.26
		Sun *et al.* [52]	69.00
CK+	FER2013	SACNN-ALSTMs(ours)	32.43
	JAFFE	SACNN-ALSTMs(ours)	44.13
		Xie *et al.* [51]	43.38
JAFFE	FER2013	SACNN-ALSTMs(ours)	34.65
	CK+	SACNN-ALSTMs(ours)	48.54
		Xie *et al.* [51]	49.10

model achieves better or competitive performance on the cross-database, which proves the proposed method is robust and applicable to practical applications.

The analyses about the comparative experimental results are summarized as follows:

- In the contrast experiments on three public databases, the recognition accuracies of SACNN-ALSTMs are higher than that of other deep learning-based methods, which verifies the effectiveness of the proposed method.
- The proposed model outperforms the baseline model by jointly combining SACNN and ALSTMs, which demonstrates that the deep geometric features of facial landmarks are conducive to improve the discriminative power of hybrid features.
- The improved recognition performance through ALSTMs proves the availability of the proposed deep geometric feature descriptor and the landmark grouping strategy. The ALSTMs can capture the dependencies of landmark groups and the attention mechanism helps the model focus on the expression-related features.
- The proposed model achieves promising performance on the FER2013 database which has complex backgrounds and head pose variations. The results suggest that our method has the ability to handle different variations in practical applications.

It is worth noting that there are still some potential threats that may reduce the recognition accuracy of the proposed method. As facial landmark points are used to extract the deep geometric feature through ALSTMs, the SACNN-ALSTMs network relies on the robust face detection and facial landmark localization. In addition, the introduction of ALSTMs leads to an increase in time and space complex.

V. PRELIMINARY APPLICATION EXPERIMENTS ON FACIAL EXPRESSION RECOGNITION

With the rapid growth of artificial intelligence and robotics, social robots promise widespread integration into human society and HRI become inevitable. However, the interaction effect is less than satisfactory because robots lack understanding of human emotional intentions. Emotion understanding becomes an essential challenge to improve the HRI performance [4], [58]. For instance, Greco *et al.* [58] propose a robotic architecture provided with emotion analysis capabilities on the basis of facial expression recognition.

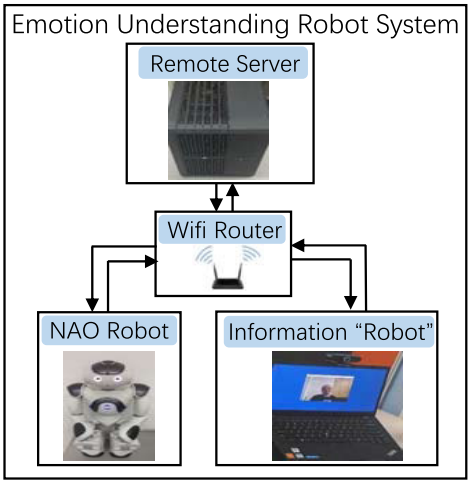


FIGURE 10. The architecture of emotion understanding robot system.

FIGURE 11. Experimental environment.

To further appraise the performance of the proposed model in a real-world application, preliminary experiments are conducted on the developing Emotion Understanding Robot System (EURS) in our lab, as shown in Fig.10. The EURS includes an NAO humanoid mobile robot with a monocular camera, an information robot, and a remote server to speed up affective computation. To facilitate the analysis of emotion recognition results, an information robot is utilized for the visualization of the recognition results. Currently, a computer is used as a substitute for the information robot. The EURS is designed to smooth human-robot interaction by realizing the robot's understanding of human emotional intention based on facial expression recognition and to make the robot respond appropriately according to the results of the emotional analysis. For instance, when the human emotion recognized in the interaction scenario is ''sadness'', the robot may start a conversation of concern and inquiry, make a gentle soothing movement, or play light music to comfort the human. Accurate emotion recognition is the premise of the subsequent robot behavior decision-making, and the proposed model based on facial expressions is a part of the preliminary work of EURS.

The experimental environment is shown in Fig.11. The NAO robot uses its monocular camera to capture the facial expressions during the human-robot interaction, and the face

TABLE 14. Comparison of Dynamic Emotional Recognition(%).

Emotion	Baseline model	SACNN-ALSTMs
Angry	72.00	76.00
Disgust	34.00	46.00
Fear	44.00	66.00
Happy	92.00	98.00
Sad	74.00	82.00
Surprise	60.00	76.00
Neutral	70.00	78.00
Average	63.71	74.57

data are uploaded to the remote server for real-time analysis. The emotion understanding results are fed back to the NAO robot for appropriate responses and sent to an information robot for visualization. Facial samples are obtained from 10 volunteers in our laboratory, including four females and six males. In the experiment, each volunteer makes seven basic expressions (i.e., happy, surprise, fear, anger, disgust, sad, and neutral), each of which is sampled for 5 times. A majority voting is adopted to validate whether the facial expressions are rightly posed. Each image is independently voted by 10 annotators, and the image is only accepted when the experimenter obtained fully recognizable expression. A total of 350 samples are collected for the preliminary application experiments on FER. As listed in Table 14, the recognition result achieves an average accuracy of 74.57%. The inference time for one expression image is about 2.8ms, which meets the requirement for the emotion recognition of EURS.

VI. CONCLUSION AND FUTURE WORK

This article presents a SACNN-ALSTMs network to extract the hybrid feature for facial expression recognition. The SACNN-ALSTMs can learn the relative geometric position dependencies of facial landmark points and extract more discriminative facial features for FER.

- The SACNN is employed to extract pixel-level features from facial pixels and the ALSTMs is utilized to explore the potentials of facial landmarks. By jointly combining the SACNN and ALSTMs, the proposed model can be more discriminative to different expressions.
- A deep geometric feature descriptor is proposed to characterize the relative geometric position relationship of facial landmarks. The geometric feature is combined with the pixel-level feature to improve the discriminative power of the facial feature.
- A landmark grouping strategy is introduced to group the facial landmarks into seven groups to capture the local geometric features. An attention module is implemented to reweigh the local geometric features which are concatenated to obtain the holistic geometric features.

The presented method is evaluated on three publicly available databases, FER2013, CK+, and JAFFE. Experimental results demonstrate the effectiveness of the proposed model.

The proposed approach shows good performance for the datasets with frontal faces and limited head deflection. However, variant head poses and occlusions are two common situations in the real world, which may directly lead to the failure detection of facial landmarks. In the future work, we will focus on exploring methods for pose and occlusion FER without landmarks, as the proposed model relies on robust face detection and facial landmark localization. Also, we plan to optimize the proposed network structure to reduce the complexity for the deployment in practical applications. In addition, the increasing number of video-based databases motivates us to develop networks for expression classification using multi-modal feature fusion.

REFERENCES

[1] J. Chen, Z. Chen, Z. Chi, and H. Fu, "Facial expression recognition in video with multiple feature fusion," *IEEE Trans. Affect. Comput.*, vol. 9, no. 1, pp. 38–50, Jan. 2018.

[2] C.-M. Kuo, S.-H. Lai, and M. Sarkis, "A compact deep learning model for robust facial expression recognition," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. Workshops (CVPRW)*, Jun. 2018, pp. 2121–2129.

[3] P. Rodriguez, G. Cucurull, J. Gonàlez, J. M. Gonfaus, K. Nasrollahi, T. B. Moeslund, and F. X. Roca, "Deep pain: Exploiting long short-term memory networks for facial expression classification," *IEEE Trans. Cybern.*, early access, Feb. 9, 2017, doi: 10.1109/TCYB.2017.2662199.

[4] L. Chen, M. Zhou, W. Su, M. Wu, J. She, and K. Hirota, "Softmax regression based deep sparse autoencoder network for facial emotion recognition in human-robot interaction," *Inf. Sci.*, vol. 428, pp. 49–61, Feb. 2018.

[5] P. Ekman, W. V. Friesen, and J. C. Hager, *Facial Action Coding System—The Manual on CD-ROM*. Salt Lake City, UT, USA: A Human Face, 2002.

[6] C. Shan, S. Gong, and P. W. McOwan, "Facial expression recognition based on local binary patterns: A comprehensive study," *Image Vis. Comput.*, vol. 27, no. 6, pp. 803–816, May 2009.

[7] J. Chen, T. Takiguchi, and Y. Ariki, "Rotation-reversal invariant HOG cascade for facial expression recognition," *Signal, Image Video Process.*, vol. 11, no. 8, pp. 1485–1492, May 2017.

[8] X. Xu, C. Quan, and F. Ren, "Facial expression recognition based on Gabor wavelet transform and histogram of oriented gradients," in *Proc. IEEE Int. Conf. Mechatronics Autom. (ICMA)*, Aug. 2015, pp. 2117–2122.

[9] F. Ren and Z. Huang, "Facial expression recognition based on AAM-SIFT and adaptive regional weighting," *IEEJ Trans. Electr. Electron. Eng.*, vol. 10, no. 6, pp. 713–722, Sep. 2015.

[10] Y. Li, J. Zeng, S. Shan, and X. Chen, "Occlusion aware facial expression recognition using CNN with attention mechanism," *IEEE Trans. Image Process.*, vol. 28, no. 5, pp. 2439–2450, May 2019.

[11] A. R. Kurup, M. Ajith, and M. M. Ramón, "Semi-supervised facial expression recognition using reduced spatial features and deep belief networks," *Neurocomputing*, vol. 367, pp. 188–197, Nov. 2019.

[12] F. Zhang, T. Zhang, Q. Mao, and C. Xu, "Joint pose and expression modeling for facial expression recognition," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2018, pp. 3359–3368.

[13] S. Li and W. Deng, "Deep facial expression recognition: A survey," *IEEE Trans. Affect. Comput.*, early access, Mar. 17, 2020, doi: 10.1109/TAFFC.2020.2981446.

[14] B. Amos, B. Ludwiczuk, and M. Satyanarayanan, "OpenFace: A general purpose face recognition library with mobile applications," School Comput. Sci., CMU, Pittsburgh, PA, USA, Tech. Rep. CMU-CS-16-118, 2016.

[15] P. Viola and M. Jones, "Rapid object detection using a boosted cascade of simple features," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, vol. 1, Dec. 2001, pp. I-511–I-518.

[16] K. Zhang, Z. Zhang, Z. Li, and Y. Qiao, "Joint face detection and alignment using multitask cascaded convolutional networks," *IEEE Signal Process. Lett.*, vol. 23, no. 10, pp. 1499–1503, Oct. 2016.

[17] V. S. Avani, S. G. Shaila, and A. Vadivel, "Geometrical features of lips using the properties of parabola for recognizing facial expression," *Cogn. Neurodyn.*, pp. 1–19, Oct. 2020, doi: 10.1007/s11571-020-09638-x.

[18] V. S. Avani, S. G. Shaila, and A. Vadivel, "Interval graph of facial regions with common intersection salient points for identifying and classifying facial expression," *Multimedia Tools Appl.*, vol. 80, pp. 3367–3390, Sep. 2020.

[19] A. Swaminathan, A. Vadivel, and M. Arock, "FERCE: Facial expression recognition for combined emotions using FERCE algorithm," *IETE J. Res.*, pp. 1–16, May 2020, doi: 10.1080/03772063.2020.1756471.
[20] B. Fasel and J. Luettin, "Automatic facial expression analysis: A survey," *Pattern Recognit.*, vol. 36, no. 1, pp. 259–275, Jan. 2003.
[21] P. Liu, S. Han, Z. Meng, and Y. Tong, "Facial expression recognition via a boosted deep belief network," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2014, pp. 1805–1812.
[22] D. K. Jain, P. Shamsolmoali, and P. Sehdev, "Extended deep neural network for facial emotion recognition," *Pattern Recognit. Lett.*, vol. 120, pp. 69–74, Apr. 2019.
[23] H.-W. Ng, V. D. Nguyen, V. Vonikakis, and S. Winkler, "Deep learning for emotion recognition on small datasets using transfer learning," in *Proc. ACM Int. Conf. Multimodal Interact. (ICMI)*, Nov. 2015, pp. 443–449.
[24] N. Kumaran, A. Vadivel, and S. S. Kumar, "Recognition of human actions using CNN-GWO: A novel modeling of CNN for enhancement of classification performance," *Multimedia Tools Appl.*, vol. 77, no. 18, pp. 23115–23147, Jan. 2018.
[25] A. Mollahosseini, D. Chan, and M. H. Mahoor, "Going deeper in facial expression recognition using deep neural networks," in *Proc. IEEE Winter Conf. Appl. Comput. Vis. (WACV)*, Mar. 2016, pp. 1–10.
[26] O. Arriaga, M. Valdenegro-Toro, and P. Plöger, "Real-time convolutional neural networks for emotion and gender classification," Oct. 2017, *arXiv:1710.07557*. [Online]. Available: http://arxiv.org/abs/1710.07557
[27] H. Jung, S. Lee, J. Yim, S. Park, and J. Kim, "Joint fine-tuning in deep neural networks for facial expression recognition," in *Proc. IEEE Int. Conf. Comput. Vis. (ICCV)*, Dec. 2015, pp. 2983–2991.
[28] Z. Yu and C. Zhang, "Image based static facial expression recognition with multiple deep network learning," in *Proc. ACM Int. Conf. Multimodal Interact. (ICMI)*, Nov. 2015, pp. 435–442.
[29] G. Levi and T. Hassner, "Emotion recognition in the wild via convolutional neural networks and mapped binary patterns," in *Proc. ACM Int. Conf. Multimodal Interact.*, Nov. 2015, pp. 503–510.
[30] T. Connie, M. Al-Shabi, W. P. Cheah, and M. Goh, "Facial expression recognition using a hybrid CNN–sift aggregator," in *Proc. Int. Workshop Multi-Disciplinary Trends Artif. Intell. (MIWAI)*. Springer, Oct. 2017, pp. 139–149.
[31] Z. Yu, G. Liu, Q. Liu, and J. Deng, "Spatio-temporal convolutional features with nested lstm for facial expression recognition," *Neurocomputing*, vol. 317, pp. 50–57, Nov. 2018.
[32] D. Liang, H. Liang, Z. Yu, and Y. Zhang, "Deep convolutional BiLSTM fusion network for facial expression recognition," *Vis. Comput.*, vol. 36, no. 3, pp. 499–508, Mar. 2020.
[33] J. Hu, L. Shen, and G. Sun, "Squeeze-and-excitation networks," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2018, pp. 7132–7141.
[34] S. Woo, J. Park, J.-Y. Lee, and I. S. Kweon, "CBAM: Convolutional block attention module," in *Proc. Eur. Conf. Comput. Vis. (ECCV)*, Sep. 2018, pp. 3–19.
[35] S. L. Happy and A. Routray, "Automatic facial expression recognition using features of salient facial patches," *IEEE Trans. Affect. Comput.*, vol. 6, no. 1, pp. 1–12, Jan. 2015.
[36] L. Zhong, Q. Liu, P. Yang, J. Huang, and D. N. Metaxas, "Learning multiscale active facial patches for expression analysis," *IEEE Trans. Cybern.*, vol. 45, no. 8, pp. 1499–1510, Aug. 2015.
[37] A. Majumder, L. Behera, and V. K. Subramanian, "Automatic facial expression recognition system using deep network-based data fusion," *IEEE Trans. Cybern.*, vol. 48, no. 1, pp. 103–114, Jan. 2018.
[38] Y. Gan, J. Chen, Z. Yang, and L. Xu, "Multiple attention network for facial expression recognition," *IEEE Access*, vol. 8, pp. 7383–7393, Jan. 2020.
[39] H. Zhang, W. Su, and Z. Wang, "Weakly supervised local-global attention network for facial expression recognition," *IEEE Access*, vol. 8, pp. 37976–37987, Feb. 2020.
[40] Y. Li, J. Zeng, S. Shan, and X. Chen, "Patch-gated CNN for occlusion-aware facial expression recognition," in *Proc. 24th Int. Conf. Pattern Recognit. (ICPR)*, Aug. 2018, pp. 2209–2214.
[41] K. Wang, X. Peng, J. Yang, D. Meng, and Y. Qiao, "Region attention networks for pose and occlusion robust facial expression recognition," *IEEE Trans. Image Process.*, vol. 29, pp. 4057–4069, Jan. 2020.
[42] A. R. Shahid, S. Khan, and H. Yan, "Contour and region harmonic features for sub-local facial expression recognition," *J. Vis. Commun. Image Represent.*, vol. 73, Nov. 2020, Art. no. 102949.
[43] I. J. Goodfellow *et al.*, "Challenges in representation learning: A report on three machine learning contests," *Neural Netw.*, vol. 64, pp. 59–63, Apr. 2015.
[44] P. Lucey, J. F. Cohn, T. Kanade, J. Saragih, Z. Ambadar, and I. Matthews, "The extended Cohn-Kanade dataset (CK+): A complete dataset for action unit and emotion specified expression," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. Workshop. (CVPRW)*, Jun. 2010, pp. 94–101.
[45] M. Lyons, S. Akamatsu, M. Kamachi, and J. Gyoba, "Coding facial expressions with Gabor wavelets," in *Proc. 3rd Int. Conf. Autom. Face Gesture Recognit.*, Apr. 1998, pp. 200–205.
[46] Y. Tang, "Deep learning using linear support vector machines," Jun. 2013, *arXiv:1306.0239*. [Online]. Available: http://arxiv.org/abs/1306.0239
[47] T. Devries, K. Biswaranjan, and G. W. Taylor, "Multi-task learning of facial landmarks and expression," in *Proc. Can. Conf. Comput. Robot Vis. (CRV)*, May 2014, pp. 98–103.
[48] J. Jeon, J.-C. Park, Y. Jo, C. Nam, K.-H. Bae, Y. Hwang, and D.-S. Kim, "A real-time facial expression recognizer using deep neural network," in *Proc. 10th Int. Conf. Ubiquitous Inf. Manage. Commun. (IMCOM)*, Jan. 2016, pp. 1–4.
[49] Y. Guo, D. Tao, J. Yu, H. Xiong, Y. Li, and D. Tao, "Deep neural networks with relativity learning for facial expression recognition," in *Proc. IEEE Int. Conf. Multimedia Expo Workshops (ICMEW)*, Jul. 2016, pp. 1–6.
[50] S. Munasinghe, C. Fookes, and S. Sridharan, "Deep features-based expression-invariant tied factor analysis for emotion recognition," in *Proc. IEEE Int. Joint Conf. Biometrics (IJCB)*, Oct. 2017, pp. 546–554.
[51] S. Xie, H. Hu, and Y. Wu, "Deep multi-path convolutional neural network joint with salient region attention for facial expression recognition," *Pattern Recognit.*, vol. 92, pp. 177–191, Aug. 2019.
[52] X. Sun, S. Zheng, and H. Fu, "ROI-attention vectorized CNN model for static facial expression recognition," *IEEE Access*, vol. 8, pp. 7183–7194, Jan. 2020.
[53] S. Taheri, Q. Qiu, and R. Chellappa, "Structure-preserving sparse decomposition for facial expression analysis," *IEEE Trans. Image Process.*, vol. 23, no. 8, pp. 3590–3603, Aug. 2014.
[54] A. T. Lopes, E. de Aguiar, A. F. De Souza, and T. Oliveira-Santos, "Facial expression recognition with convolutional neural networks: Coping with few data and the training sample order," *Pattern Recognit.*, vol. 61, pp. 610–628, Jan. 2017.
[55] T. Zhang, W. Zheng, Z. Cui, Y. Zong, and Y. Li, "Spatial–temporal recurrent neural network for emotion recognition," *IEEE Trans. Cybern.*, vol. 49, no. 3, pp. 839–847, Jan. 2018.
[56] J. Cai, Z. Meng, A. S. Khan, Z. Li, J. O'Reilly, and Y. Tong, "Island loss for learning discriminative features in facial expression recognition," in *Proc. 13th IEEE Int. Conf. Autom. Face Gesture Recognit.*, May 2018, pp. 302–309.
[57] K. Li, Y. Jin, M. W. Akram, R. Han, and J. Chen, "Facial expression recognition with convolutional neural networks via a new face cropping and rotation strategy," *Vis. Comput.*, vol. 36, no. 2, pp. 391–404, Feb. 2020.
[58] A. Greco, A. Roberto, A. Saggese, M. Vento, and V. Vigilante, "Emotion analysis from faces for social robotics," in *Proc. IEEE Int. Conf. Syst., Man Cybern. (SMC)*, Oct. 2019, pp. 358–364.

CHANG LIU received the M.E. degree from the School of Electrical Engineering and Automation, Tianjin Polytechnic University, Tianjin, China, in 2016. He is currently pursuing the Ph.D. degree with the School of Automation, Beijing Institute of Technology, Beijing, China. His research interests include affective computing, computational intelligence, and deep learning.

KAORU HIROTA (Life Member, IEEE) received the Dr.E. degree from the Tokyo Institute of Technology, Tokyo, Japan, in 1979.

He is currently a Professor with the School of Automation, Beijing Institute of Technology, Beijing, China. His current research interests include fuzzy systems, intelligent robots, and image understanding.

Dr. Hirota was a Fellow and an Experienced President of the International Fuzzy Systems Association, and a President of the Japan Society for Fuzzy Theory and Systems. He was a recipient of the Banki Donat Medal, the Henri Coanda Medal, the Grigore MOISIL Award, the SOFT Best Paper Award, and the Acoustical Society of Japan Best Paper Award. He has organized more than ten international conferences/symposiums as the founding/general/program chair. He is a Chief Editor of the *Journal of Advanced Computational Intelligence and Intelligent Informatics*.

JUNJIE MA received the B.E. degree from the Taiyuan University of Technology, Taiyuan, China, in 2012, and the Ph.D. degree from the Beijing Institute of Technology, Beijing, China, in 2020. He was founded by the China Scholarship Council and as an Internship Student with Nanyang Technological University, from 2017 to 2019. He is currently a Post Ph.D. Researcher with the Department of Computer Science and Technology, Tsinghua University, Beijing, China. His research interests include computer vision and deep learning.

ZHIYANG JIA (Member, IEEE) received the B.E. degree from the Department of Electrical Engineering and Automation, Northwestern Polytechnical University, Xi'an, China, in 2010, the M.E. degree from the Engineering Center for Digital Community of Ministry of Education, Department of Control Science and Engineering, Beijing University of Technology, Beijing, China, in 2013, and the Ph.D. degree in electrical and computer engineering from the University of Connecticut, Storrs, CT, USA, in 2017.

He is currently an Assistant Professor with School of Automation, Beijing Institute of Technology. His research interests include the smart manufacturing, modeling, analysis, and control of production systems.

YAPING DAI received the M.S. degree from the School of Automation, Beijing University of Chemical Technology, Beijing, China, in 1990, and the Ph.D. degree from the School of Automation, Beijing Institute of Technology, Beijing, in 1993.

She is currently a Professor with the School of Automation, Beijing Institute of Technology. Her current research interests include affective computing, decision support systems, and image processing.